Inferência Estatística II

Análise Bivariada

Prof. Dr. Juliano van Melis

Parte I

Parte 1

Conteúdo

Comparação de duas médias

- Usando o R
- Dimensionamento da amostra para testes de diferenças
- Teste t student para diferença de duas médias provenientes de amostras independentes
- Teste t student para diferença de duas amostras relacionadas

Parte 1 e ½

Conteúdo

Comparação de proporções

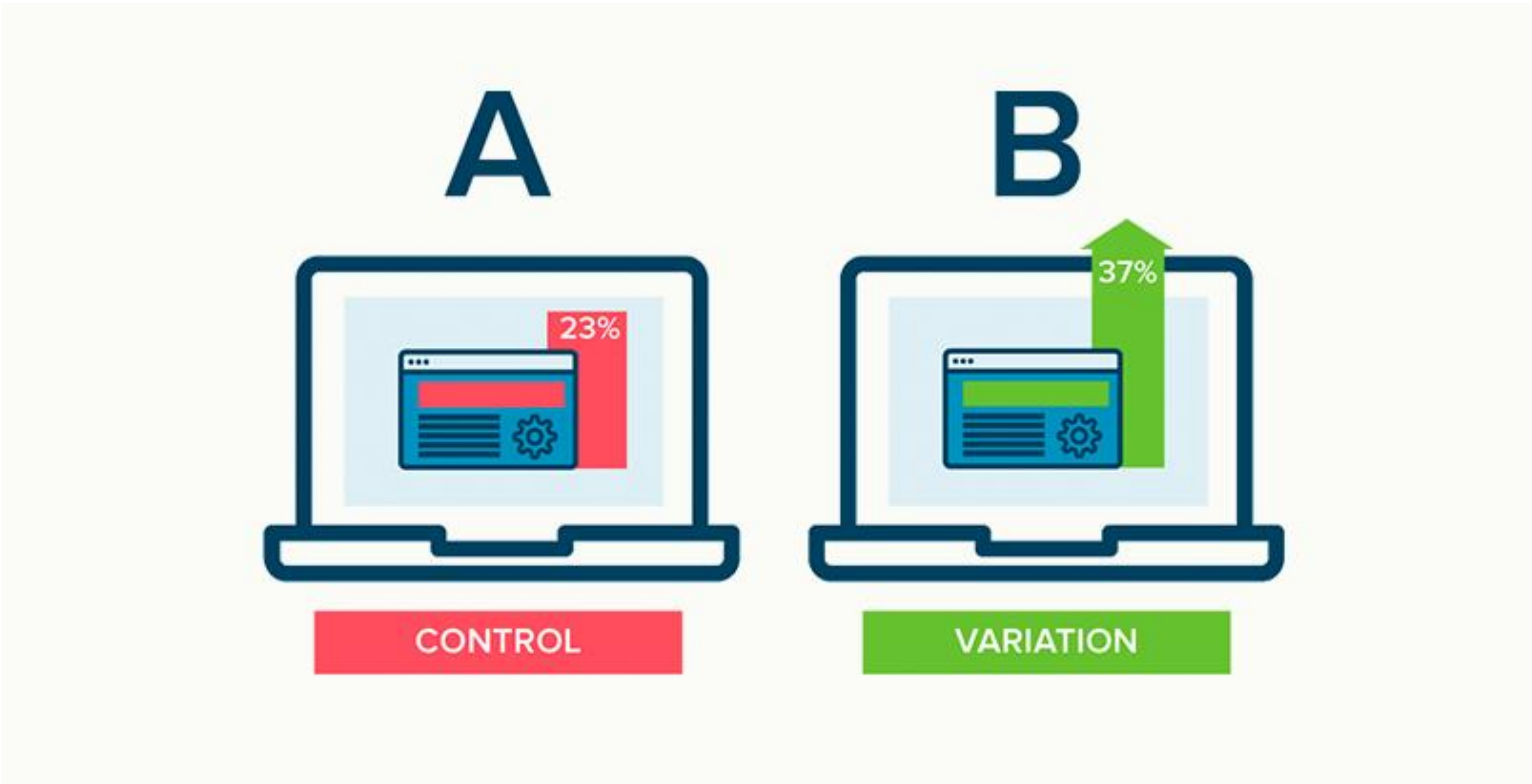
- teste Z para diferença entre duas proporções
- teste Qui-quadrado para diferença de duas proporções
- teste Qui-quadrado para Independência

Conteúdo

Parte 2

Testes não-paramétricos

- Teste dos Sinais
- Teste Mann-Witney
- Teste de Wilcoxon



Análise Bivariada

COMPARAÇÃO DE DUAS MÉDIAS

Objetivo: Comparar duas médias

Hipótese Nula: As médias são **<u>iguais</u>**

$$\boxed{\mu_A = \mu_B}$$

<u>Hipótese Alternativa</u>: As médias *<u>não</u>* são iguais

$$\boxed{\mu_A \neq \mu_B}$$

Vimos que, para o teste t para uma amostra, a estatística era definida como:

$$t = \frac{\text{média da amostra} - \text{valor nulo}}{\text{erro padrão}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Mas agora possuímos **dois** grupos amostrais, portanto podemos desenvolver a equação acima em função de **duas** amostras:

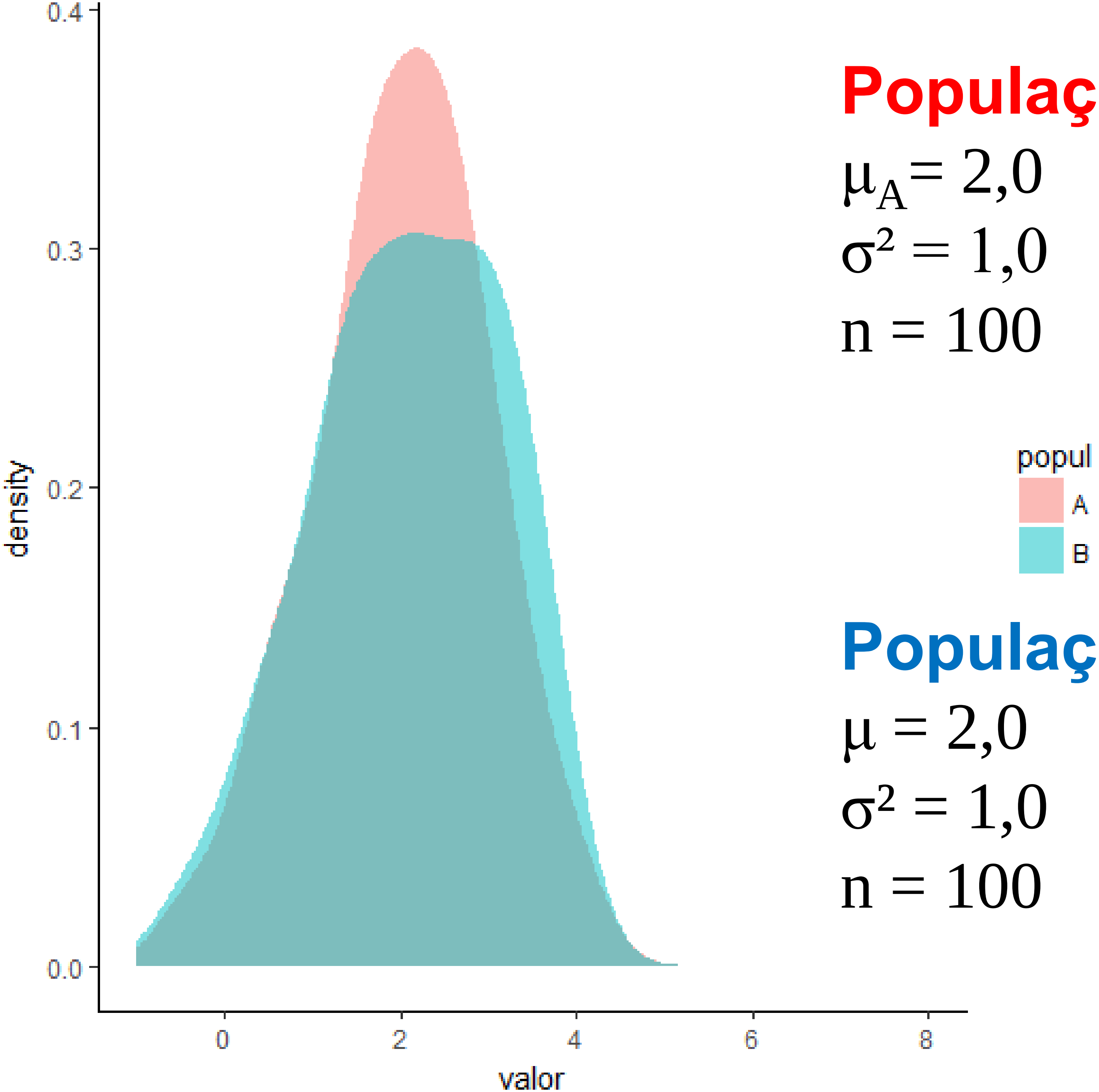
A **<u>hipótese nula</u>** deve ser representada como hipótese de **<u>igualdade</u>**.

→ Se nossa hipótese nula estiver correta, os dois grupos (amostras) foram retirados da mesma população, portanto:

$$t = \frac{(méd\ amostral_1 - méd\ amostral_2) - (méd\ populacional_1 - méd\ populacional_2)}{estimativa\ do\ erro\ padrão}$$

$$erro\ padrão\ amostral = \frac{desvio\ padrão\ amostral}{\sqrt{N}}$$

~ Distribuição t de Student



População "A"

$\mu_A = 2{,}0$

$\sigma^2 = 1{,}0$

$n = 100$

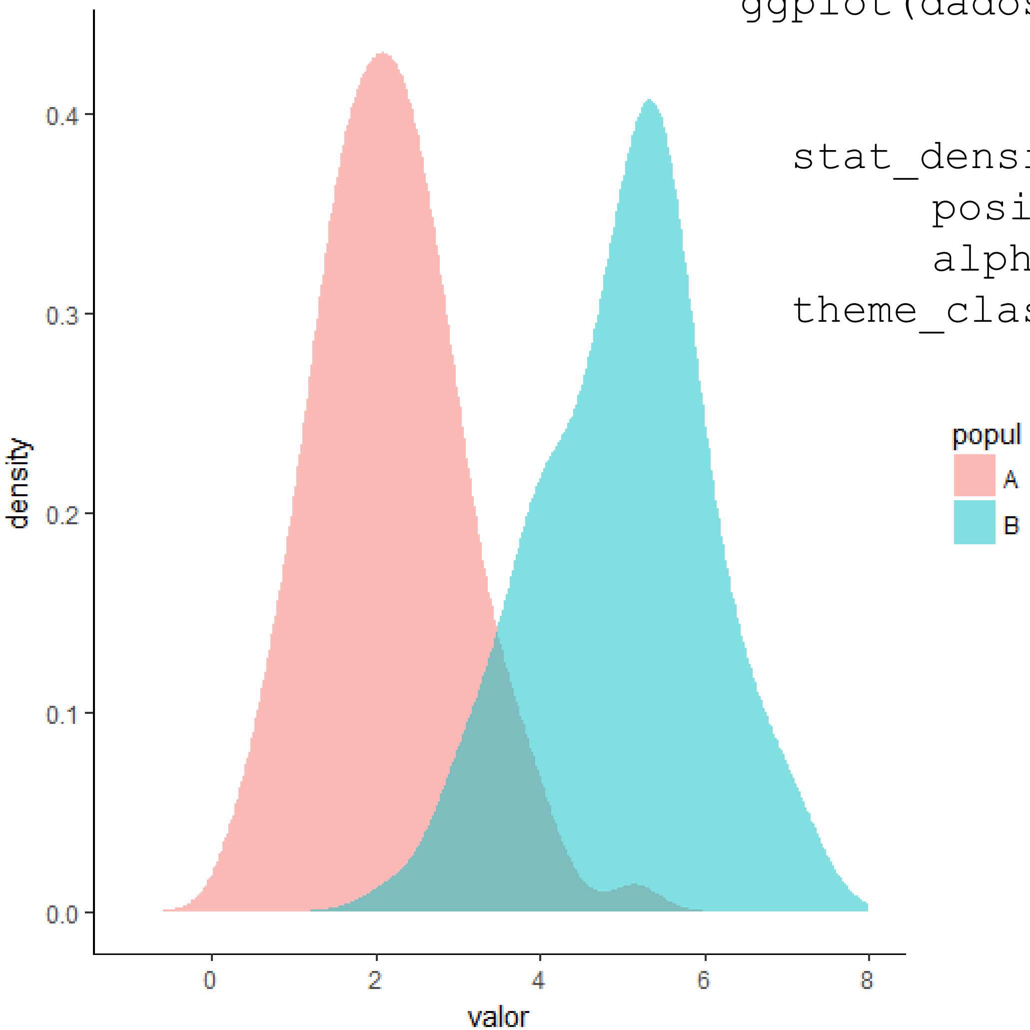
População "B"

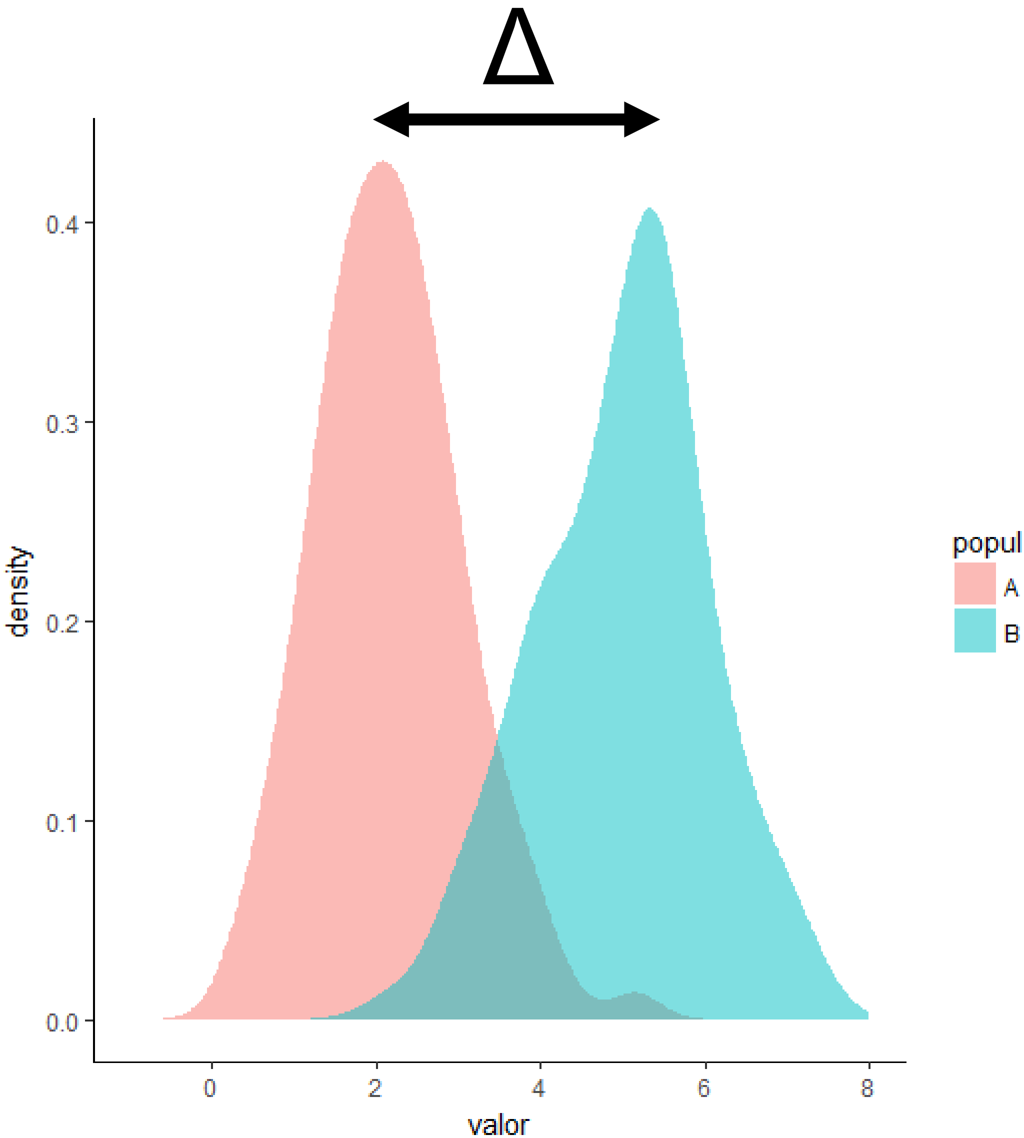
$\mu = 2{,}0$

$\sigma^2 = 1{,}0$

$n = 100$

```
dados<-data.frame(valor = c(rnorm(100, 2),
                            rnorm(100,5)),
                  popul = c(rep('A',100),
                            rep('B',100)))

require(ggplot2)

ggplot(dados, aes(x=valor,
                  group=popul,
                  fill=popul))+
  stat_density(
      position= position_dodge(width = 0),
      alpha=.5)+
  theme_classic()+xlim(c(-1,8))
```





População “A”

`rnorm(100, 2)`

N= 100

Média = 2

Desvio-Padrão = 1

População “B”

`rnorm(100, 5)`

N= 100

Média = 5

Desvio-Padrão = 1

Pressupostos do Teste

1. Distribuição normal dos dados

→ Teste de Shapiro-Wilk

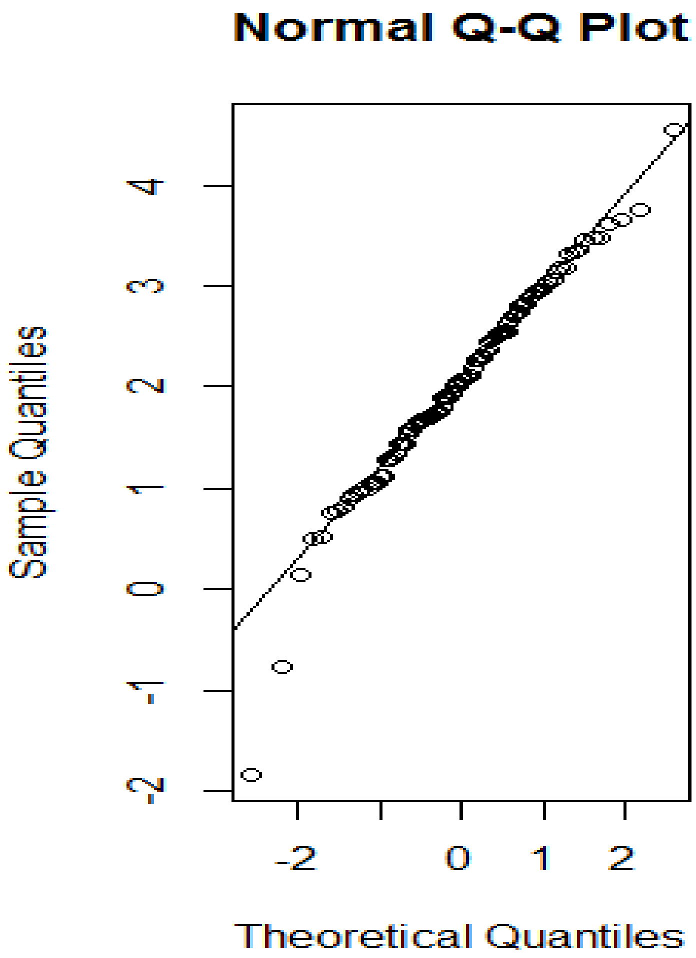
```
shapiro.test(teste_1$valor[teste_1$popul=='A'])
shapiro.test(teste_1$valor[teste_1$popul=='B'])
```

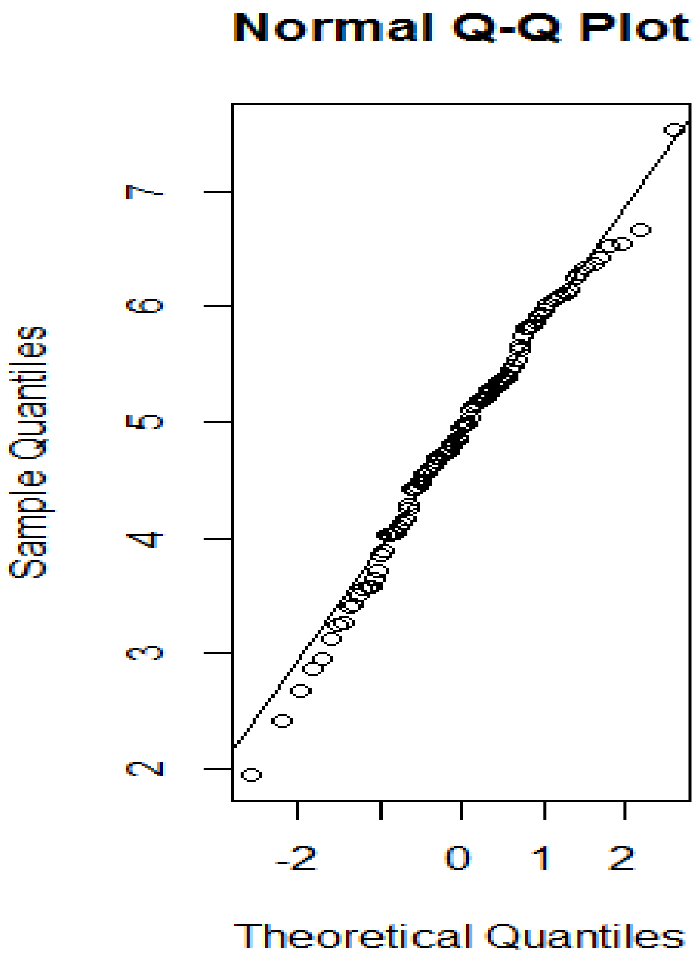
→ Gráfico Quantil-Quantil

```
qqnorm(teste_1$valor[teste_1$popul=='A'])
qqnorm(teste_1$valor[teste_1$popul=='B'])
```

Pressupostos do Teste

1. Distribuição normal dos dados





1º Passo: Estabelecer Hipótese Nula

2º Passo: Estabelecer a probabilidade de erro (Tipo I)

3º Passo: Calcular a estatística do teste

4º Passo: Concluir

1º Passo: Estabelecer Hipótese Nula
2º Passo: Estabelecer a probabilidade de erro (Tipo I)
3º Passo: Calcular a estatística do teste
4º Passo: Concluir

TESTE DE HIPÓTESES

		Situação Verdadeira	
		H0 é verdadeira	**H0 é falsa**
DECISÃO	**Rejeitar H0**	Erro tipo I (α) “falso positivo”	Poder (1 - β)
	NÃO Rejeita H0 **(H0 é aceita)**	1 - α	Erro tipo II (β) “falso negativo”

1º Passo: Estabelecer Hipótese Nula

2º Passo: Estabelecer a probabilidade de erro (Tipo I)

3º Passo: Calcular a estatística do teste

4º Passo: Concluir

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}$: média da amostra x

s_x : desvio padrão amostra x

n_1 : número de amostras x

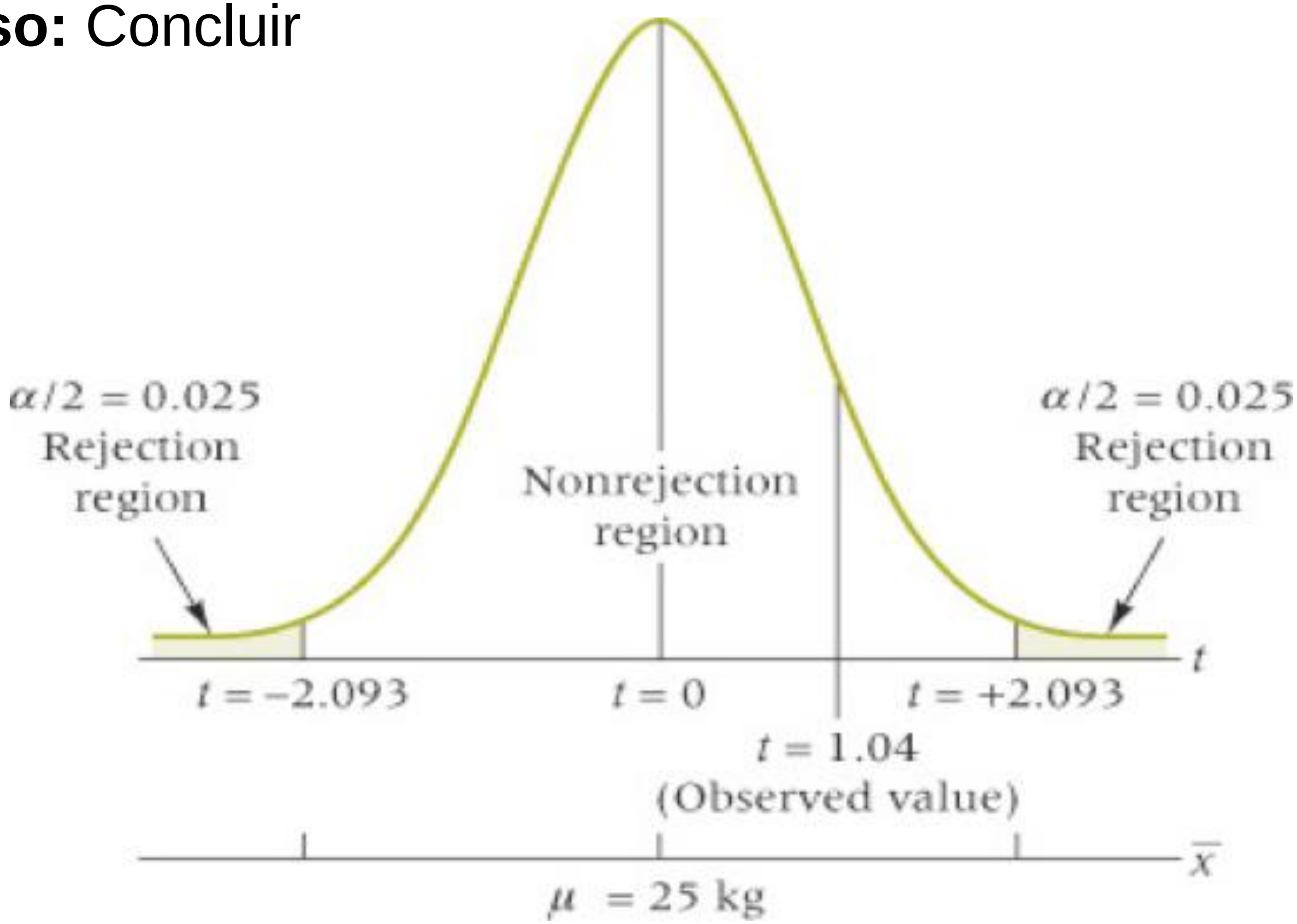
Introdução	Dimensionamento da amostra	Teste t Student Independente	Teste t Student Dependente

1º Passo: Estabelecer Hipótese Nula

2º Passo: Estabelecer a probabilidade de erro (Tipo I)

3º Passo: Calcular a estatística do teste

4º Passo: Concluir



EXERCÍCIO

Um estudo sobre regulagem de motores (mesma marca e modelo e condições) mostrou que os **57 veículos** que não passaram pela regulagem consumiram, em media, **105.32** litros de combustível, com um desvio padrão de **14.68** litros. Os **17 veículos** que passaram pela regulagem consumiram **96.82** litros de combustível com um desvio padrão de **14,26** litros. As distribuições de consumo são aproximadamente normais.

Pode-se afirmar que a regulagem reduziu o consumo de combustível (utilizar alfa de 0,05)?

EXERCÍCIO

<u>Reprovados</u>

N = **57 veículos**

Média = **105.32** litros de combustível,

Desvio padrão = **14.68** litros.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

<u>Aprovados</u>

N = 17 veículos

Média = **96.82** litros de combustível

Desvio padrão = **14,26** litros

→ **As distribuições de consumo são aproximadamente normais.**

Pode-se afirmar que a regulagem reduziu o consumo de combustível (utilizar alfa de 0,05)?

EXERCÍCIO

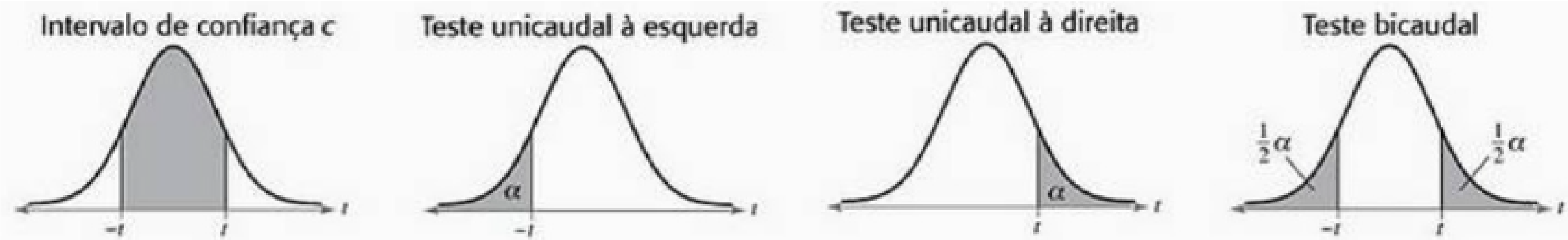


TABLE of CRITICAL VALUES for STUDENT'S *t* DISTRIBUTIONS

Column headings denote probabilities (α) ***above*** tabulated values.

d.f.	0.40	0.25	0.10	0.05	0.04	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	7.916	12.706	15.894	31.821	63.656	127.321	318.289	636.578
2	0.289	0.816	1.886	2.920	3.320	4.303	4.849	6.965	9.925	14.089	22.328	31.600
3	0.277	0.765	1.638	2.353	2.605	3.182	3.482	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.333	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.191	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	0.265	0.718	1.440	1.943	2.104	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.046	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.004	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	1.973	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	1.948	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	1.928	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	1.912	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	1.899	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	1.887	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	1.878	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	1.869	2.120	2.235	2.583	2.921	3.252	3.686	4.015

EXERCÍCIO

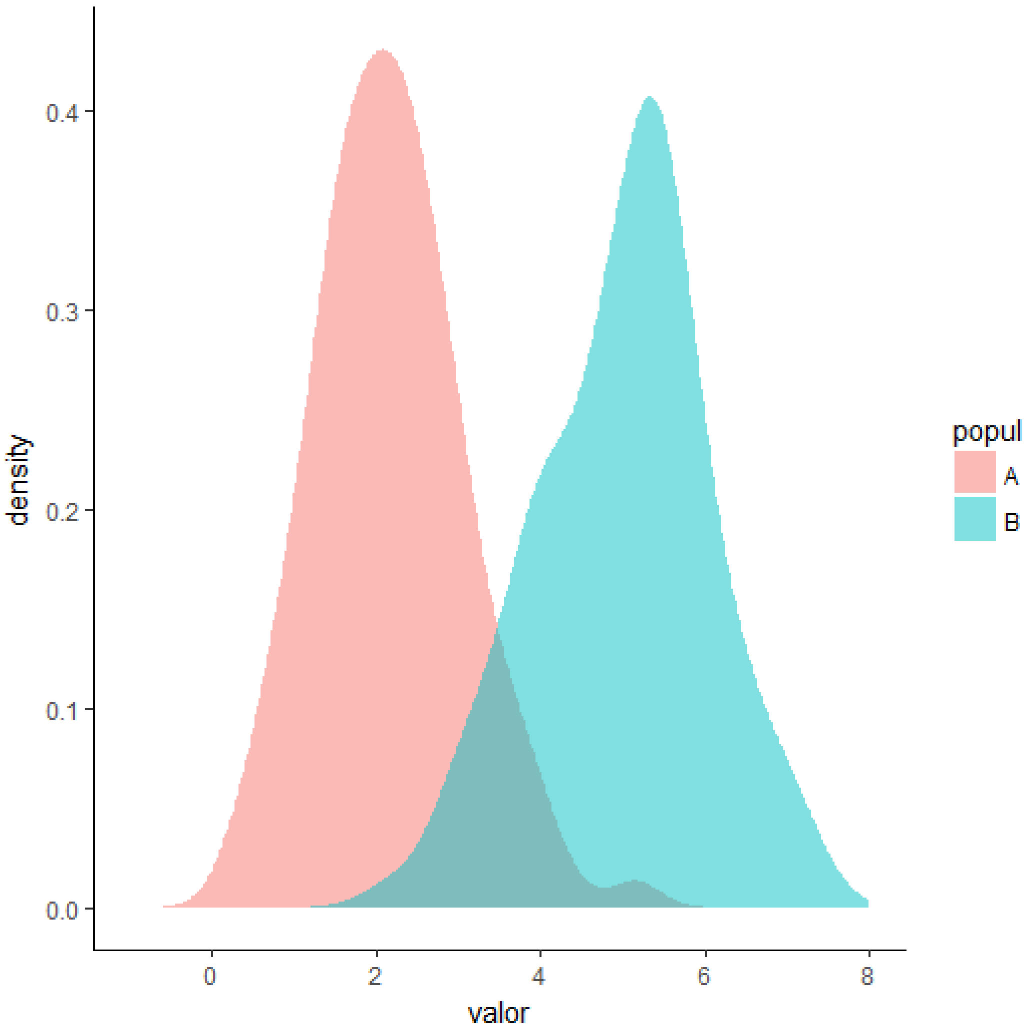
1º Ler o arquivo "teste_1.csv"

→Utilize `read.csv()`

2º Usar `t.test()`

→ Utilize `?t.test()`

3º Concluir



População “A”

`rnorm(100, 2)`

N= 100

Média = 2

Desvio-Padrão = 1

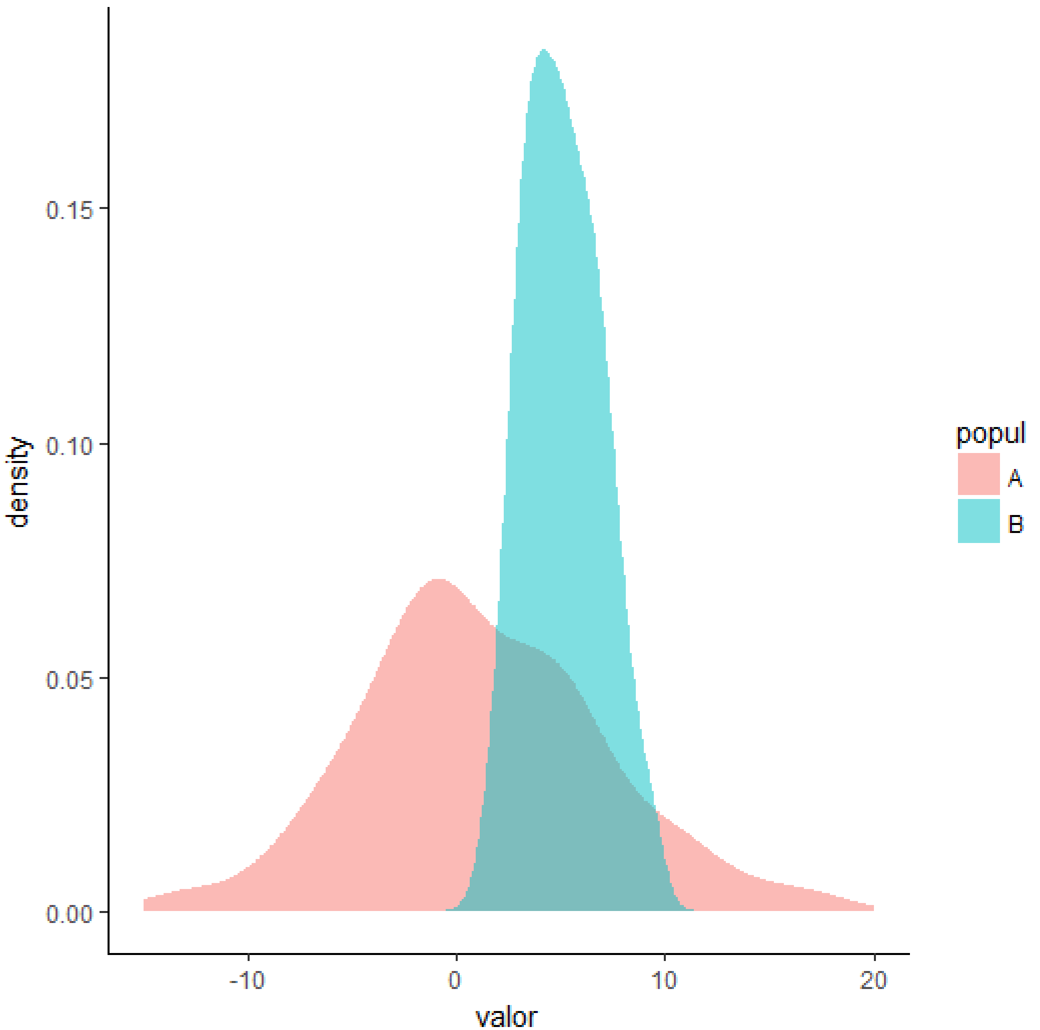
População “B”

`rnorm(100, 5)`

N= 100

Média = 5

Desvio-Padrão = 1



População “A”

`rnorm(100,2,5)`

N= 100

Média = 2

Desvio-Padrão = 5

População “B”

`rnorm(100,5,2)`

N= 100

Média = 5

Desvio-Padrão = 2

Pressupostos do Teste

1. Distribuição normal dos dados
2. Variâncias iguais (Homocedasticidade)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Usar menor grau de liberdade

g.l. : n1 – 1

g.l. : n2 – 1

Como testar Variâncias?

Hipótese nula: Variâncias são iguais

Hipótese alternativa: Variâncias não são iguais

Chamemos de S_1^2 e S_2^2 as variâncias amostrais respectivas. De (13.3) e sob a suposição de H_0 ser verdadeira, isto é $\sigma_1^2 = \sigma_2^2$, temos que

$$W = S_1^2/S_2^2 \sim F(n-1,\, m-1). \tag{13.4}$$

→Teste de Levene

```
require(car)
leveneTest(valor ~ popul, teste_2)
```

Como testar Variâncias?

→Teste de Levene

W ~ Distribuição F
gl_1= k-1 e gl_2= N-k

Definition [edit]

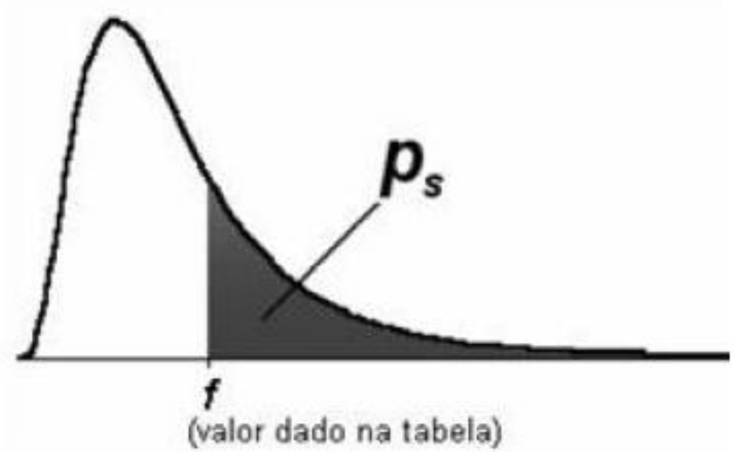
The test statistic, W, is defined as follows:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

where

- k is the number of different groups to which the sampled cases belong,
- N_i is the number of cases in the ith group,
- N is the total number of cases in all groups,
- Y_{ij} is the value of the measured variable for the jth case from the ith group,
- $Z_{ij} = \begin{cases} |Y_{ij} - \bar{Y}_{i\cdot}|, & \bar{Y}_{i\cdot} \text{ is a mean of the } i\text{-th group}, \\ |Y_{ij} - \tilde{Y}_{i\cdot}|, & \tilde{Y}_{i\cdot} \text{ is a median of the } i\text{-th group}. \end{cases}$

DISTRIBUIÇÃO F



	p	Degrees of freedom in numerator (df1)										
		1	2	3	4	5	6	7	8	12	24	1000
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.30
	0.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.2
	0.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	976.7	997.3	1017.8
	0.010	4052	4999	5404	5624	5764	5859	5928	5981	6107	6234	6363
	0.001	405312	499725	540257	562668	576496	586033	593185	597954	610352	623703	636101
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.41	39.46	39.50
	0.010	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.42	99.46	99.50
	0.001	998.38	998.84	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31

1º Ler o arquivo “teste_2.csv”

→Utilize `read.csv()`

2º Usar `t.test()`

→ Utilize `?t.test()`

3º Avalie as variâncias

4º Concluir

→ **Quando as variâncias são distintas, utilizamos o teste t de Welch**

```
t.test(valor ~ popul, teste_2, var.equal = FALSE)
```

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}}$$

Graus de Liberdade

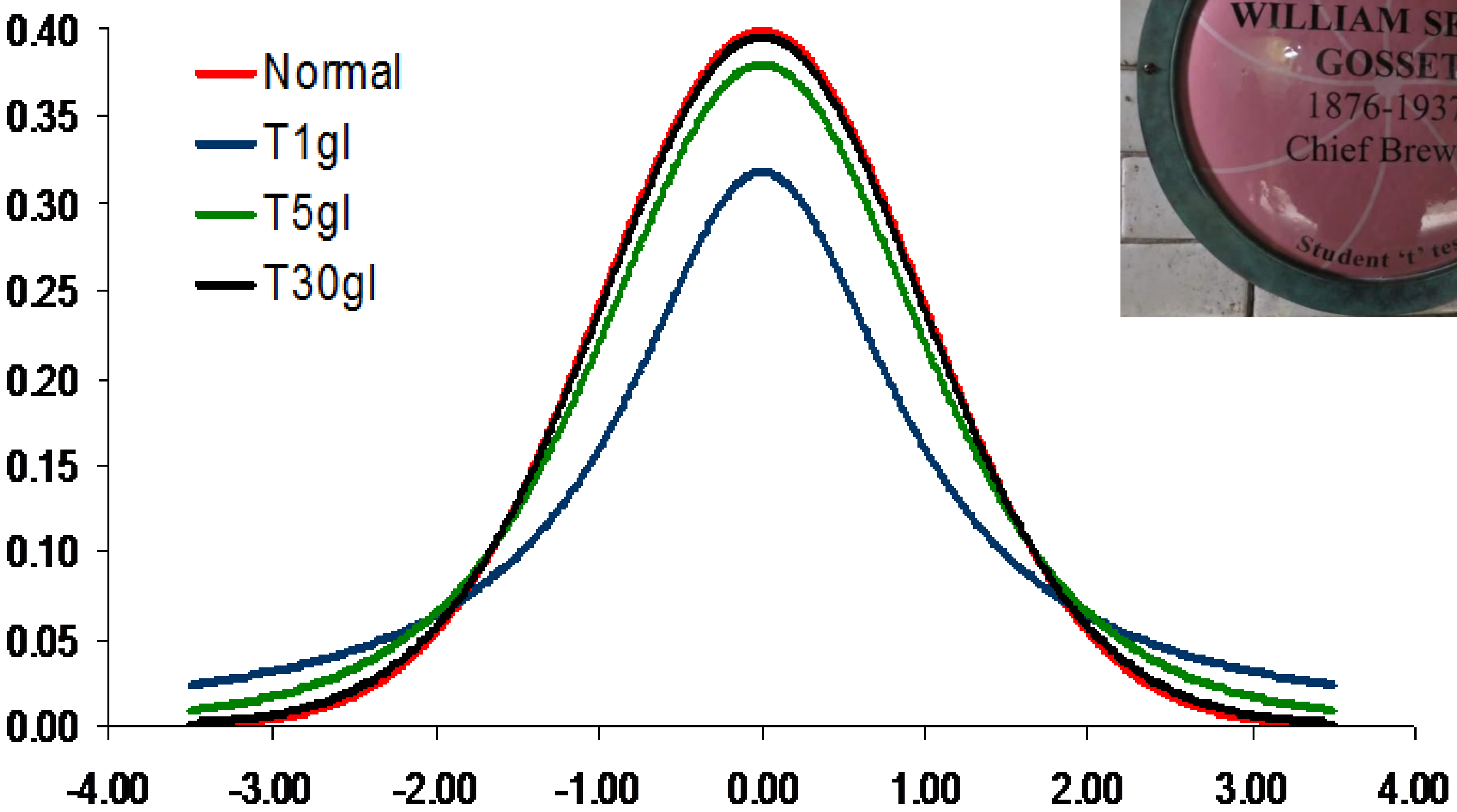
```
t.test(valor ~ popul, teste_2, var.equal = FALSE)
```

```
        Welch Two Sample t-test

data:  valor by popul
t = 3.1682, df = 35.33, p-value = 0.00316
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.452009 6.627564
sample estimates:
mean in group A mean in group B 
       25.16017        21.12038 
```

```
t.test(valor ~ popul, teste_2, var.equal = TRUE)
```

```
        Two Sample t-test

data:  valor by popul
t = 3.1682, df = 58, p-value = 0.002448
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.487347 6.592226
sample estimates:
```

Aproximação da distribuição t de Student a Normal



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\boxed{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

Erro Padrão

O ideal que o Erro Padrão seja baixo → Número amostral grande

→Tamanho do Efeito (*Effect Size*)

O mais comum é reportar a **significância** dos resultados obtidos nas pesquisas (***p-value***).

Mas é importante avaliar o **significado** (a importância prática) dos resultados de eventuais diferenças encontradas entre duas ou mais **médias** ou **variâncias**.

Testes para calcular Tamanho do Efeito

-Usando as Médias

Cohen d; Glass Δ, Hedges g, Psi Ψ

-Usando as Variâncias

Pearson r^2, Eta^2 η^2, Omega^2 Ω^2, Cohen f^2

→Tamanho do Efeito (*Effect Size*)

Always present effect sizes for primary outcomes...If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (*r* or *d*).

— L. Wilkinson and APA Task Force on Statistical Inference (1999, p. 599)

"d" de Cohen

Pode ser utilizado quando o estudo abrange duas amostras que apresentam grupos independentes e de mesmo tamanho.

→Tamanho do Efeito (*Effect Size*)

Cohen d

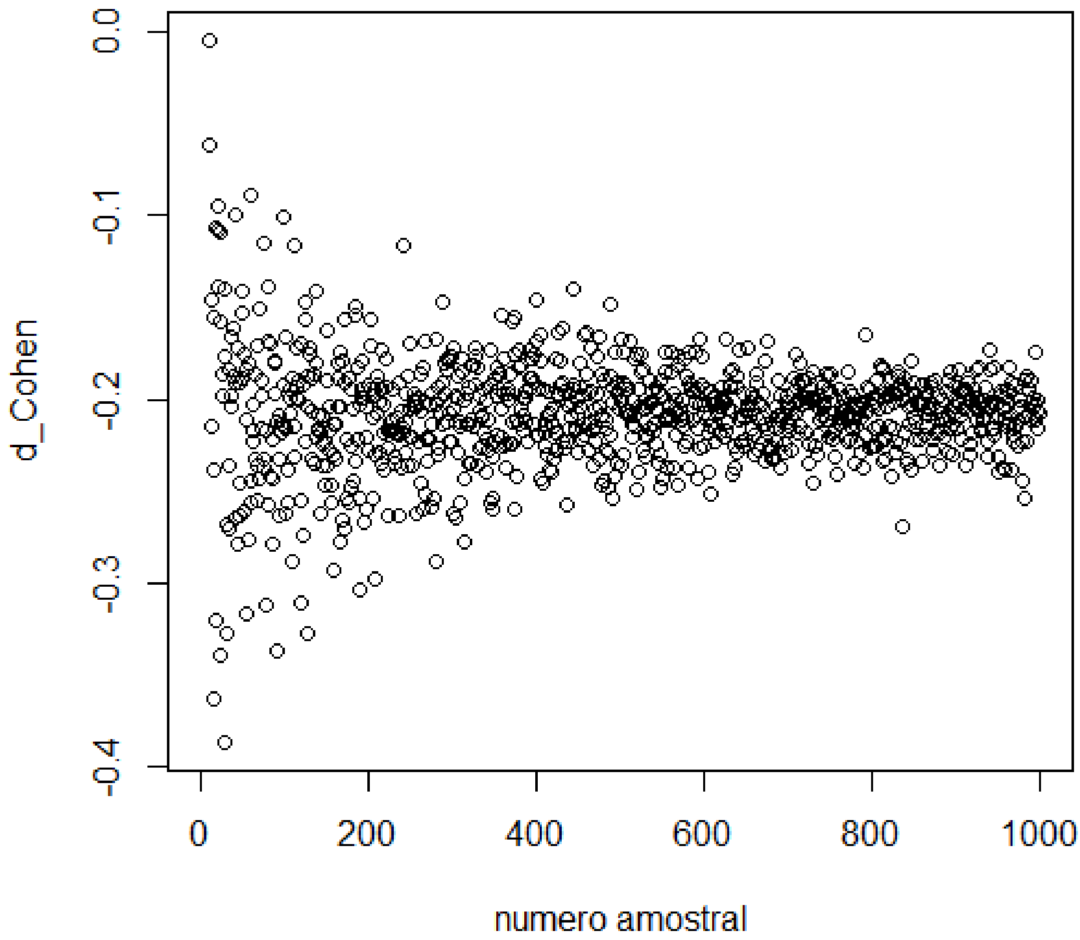
$$\mathbf{d} = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Tamanho do Efeito	d
Pequeno	0.20 – 0.30
Médio	0.40 – 0.70
Grande	≥ 0.80

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

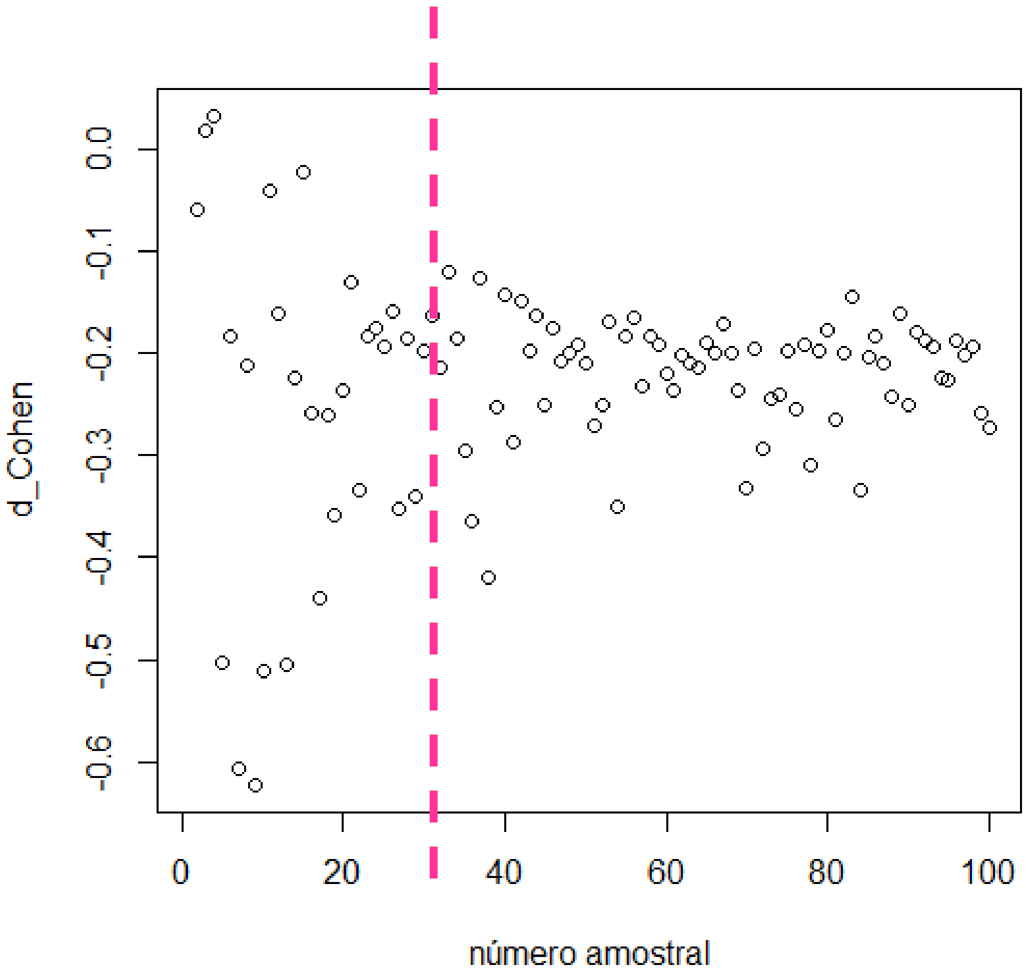
```
require(DescTools)
CohenD(x = teste_2$valor[teste_2$popul=='A'],
       y = teste_2$valor[teste_2$popul=='B'])
```

→Tamanho do Efeito (*Effect Size*)

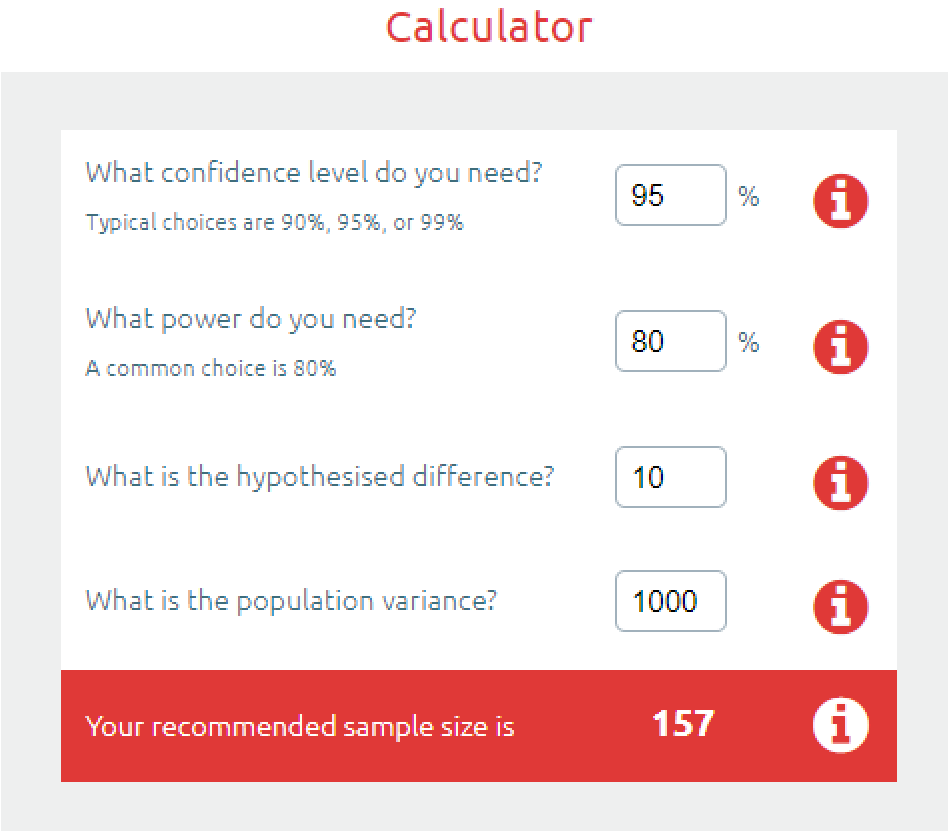


Quanto maior o tamanho amostral, mais próximo do valor do Tamanho do Efeito *real*.

→Tamanho do Efeito (*Effect Size*)



Quanto MENOR o ***Tamanho do Efeito*** indica a necessidade de um ***Tamanho Amostral*** MAIOR.



$$n = (Z_{\alpha/2}+Z_{\beta})^2 * 2 * \sigma^2 / d^2$$

$Z_{\alpha/2}$: valor Z crítico de uma distribuiçao Normal com

$\alpha/2$: metade do erro tipo I (intervalo de confiança = 1 – α)

Z_{β}: valor Z crítico com valor de

β: relativo ao erro tipo II (poder do teste = 1-β)

σ^2: Variância da *<u>população</u>*

d: Diferença esperada em detectar.

https://select-statistics.co.uk/calculators/sample-size-calculator-two-means/

1º Ler o arquivo “teste_3.csv”

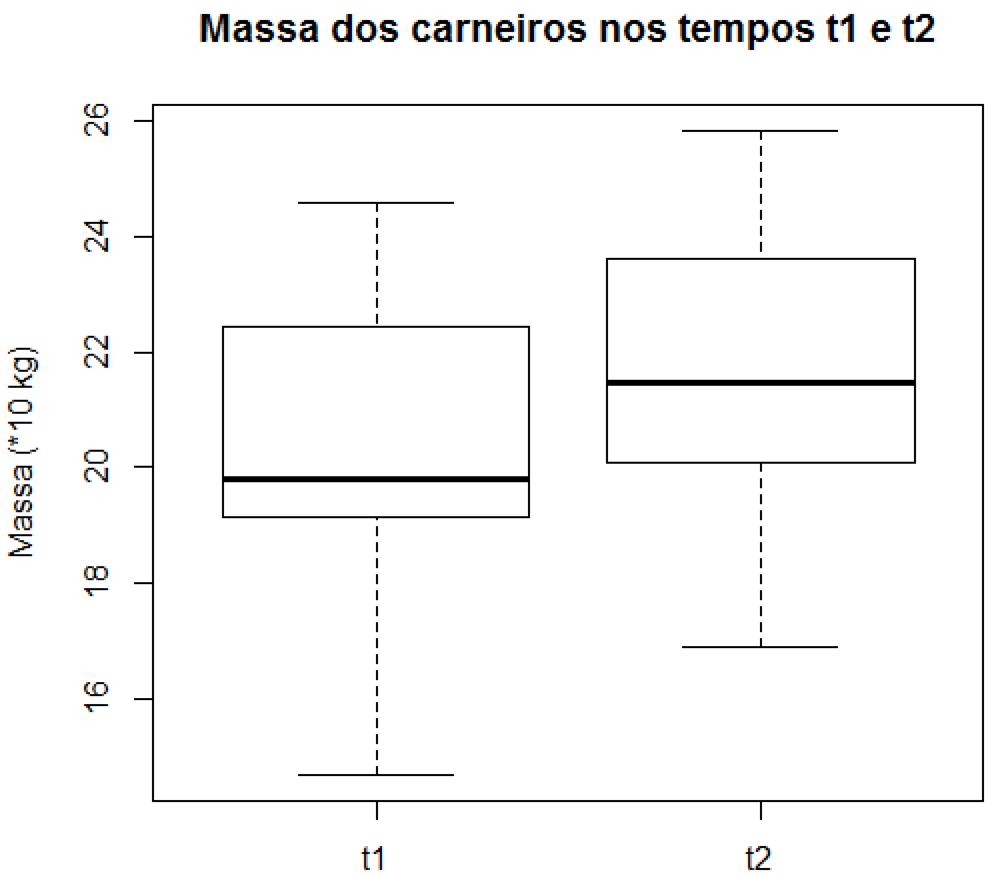
2º Explorar os dados

- Construir um *boxplot*
- Ver as variâncias das duas populações
- Testar igualdade das variâncias
- Ver as médias das duas populações
- Verificar se dados são normais

3º Efetuar um teste t (*qual deles?*)

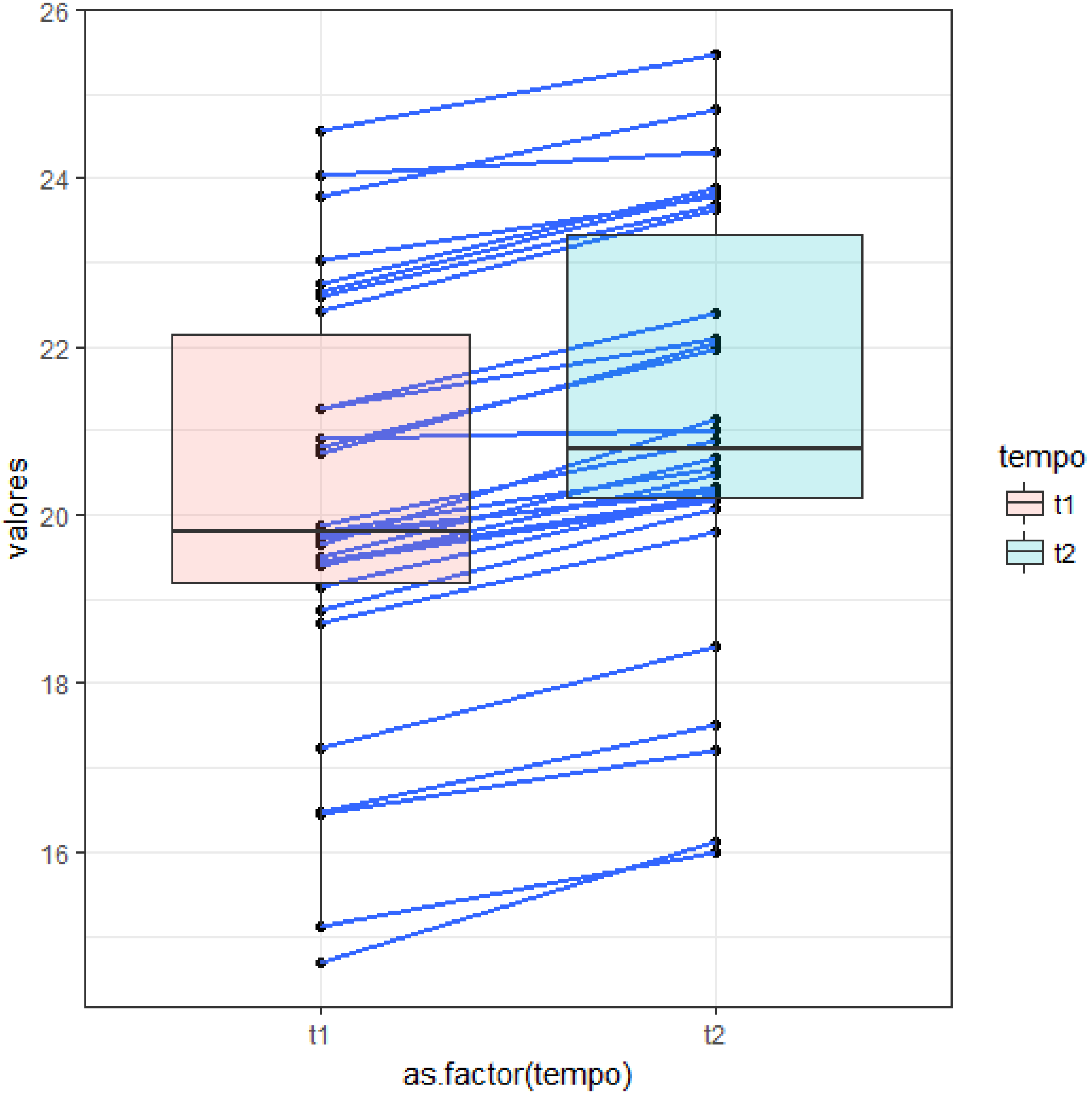
4º Calcular o Tamanho do Efeito

5º Concluir



Um estudo quer verificar o efeito de uma dieta no crescimento (em 10 kilogramas) de carneiros.
Para isso, ele mensurou a massa de 30 carneiros em dois tempos: t1 e t2.
No tempo t1 os carneiros não iniciaram a dieta e no tempo t2 depois que iniciaram a dieta.

Problema:
-São os **mesmos** carneiros mensurados antes e depois.


26
24
22
20
18
16
valores
t1
t2
as.factor(tempo)
tempo
t1
t2

Pressupostos do Teste

1. Distribuição normal dos dados
2. Variâncias iguais (Homocedasticidade)

   → Ajuste no grau de liberdade
3. As observações devem ser independentes

   → Teste com dependência

$$T = \frac{\overline{D}}{\sqrt{S_D^2 / n}}$$

D = Medida$_{depois}$ - Medida$_{antes}$

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \overline{D})^2.$$

```
t.test(valores ~ tempo, teste_4_long)
```

```
        Welch Two Sample t-test

data:  valores by tempo
t = -1.4999, df = 57.981, p-value = 0.1391
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.249186  0.322331
sample estimates:
mean in group t1 mean in group t2 
        20.13717         21.10060 
```

```
t.test(valores ~ tempo, teste_4_long,
       paired=TRUE)
```

```
        Paired t-test

data:  valores by tempo
t = -16.751, df = 29, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0810567 -0.8457982
sample estimates:
mean of the differences 
             -0.9634274 
```

EXERCÍCIOS

1. Num estudo comparativo do tempo médio de adaptação, uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatísticas	Homens	Mulheres
Médias Desvios padrões	3,2 anos 0,8 anos	3,7 anos 0,9 anos

Que conclusões você poderia tirar para a população de homens e mulheres dessa indústria? (Indique as suposições feitas para resolver o problema.)

EXERCÍCIOS

2. Uma fábrica de embalagens para produtos químicos está estudando dois processos para combater a corrosão de suas latas especiais. Para verificar o efeito dos tratamentos, foram usadas amostras cujos resultados estão no quadro abaixo (em porcentagem de corrosão eliminada). Qual seria a conclusão sobre os dois tratamentos?

Método	Amostra	Média	Desvio Padrão
A	15	48	10
B	12	52	15

EXERCÍCIOS

3. Cinco operadores de certo tipo de máquina são treinados em máquinas de duas marcas diferentes, A e B. Mediu-se o tempo que cada um deles gasta na realização de uma mesma tarefa, e os resultados estão na Tabela 13.8.

Tabela 13.8: Tempos para realização de tarefa para cinco operadores.

Operador	Marca A	Marca B
1	80	75
2	72	70
3	65	60
4	78	72
5	85	78

Com o nível de significância de 10%, poderíamos afirmar que a tarefa realizada na máquina A demora mais do que na máquina B?

FÓRMULAS

→ **quando variâncias desconhecidas e distintas**

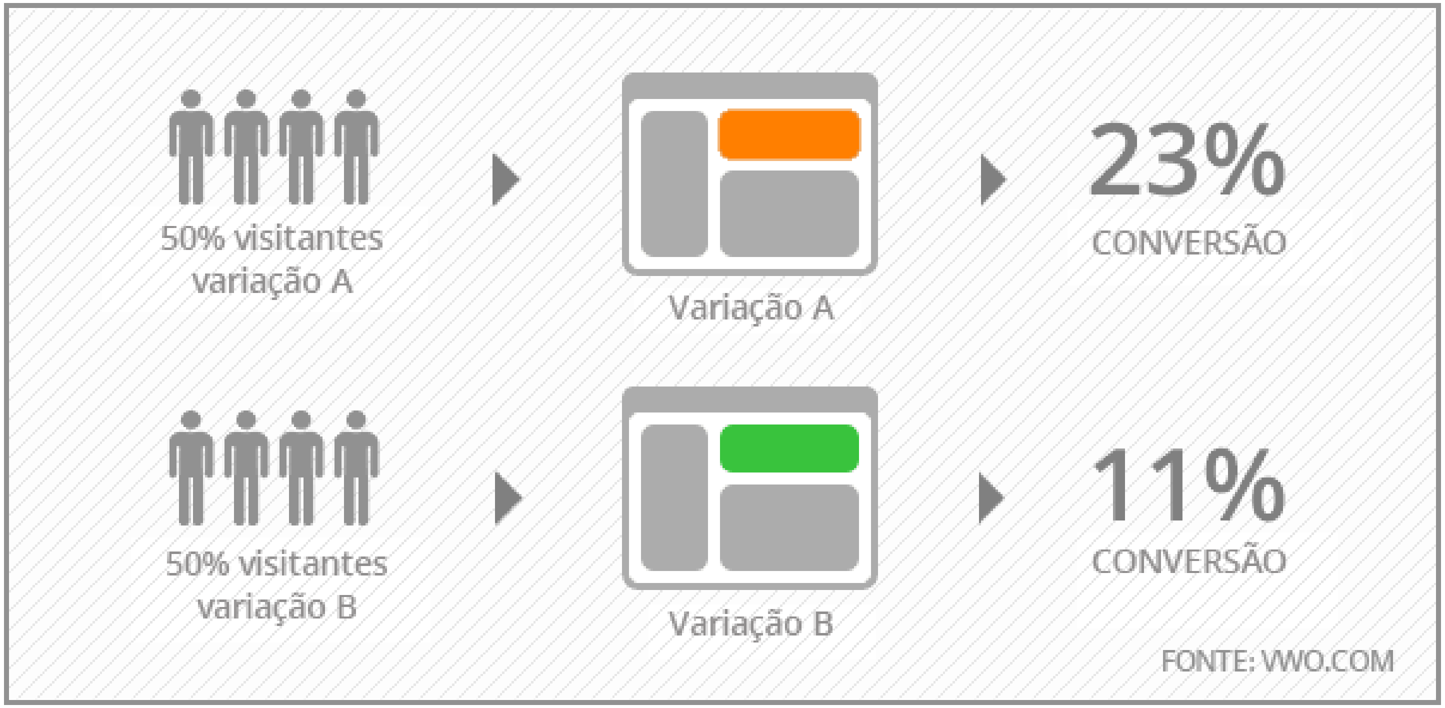
$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}}.$$

$$\nu = \frac{(A + B)^2}{A^2/(n-1) + B^2/(m-1)}$$

→ **quando amostras são pareadas**

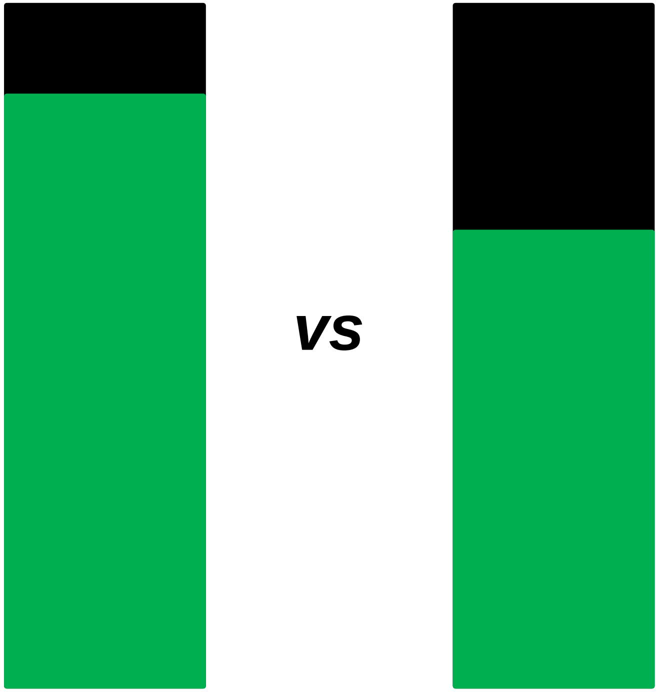
$$t = \frac{diferença}{\sqrt{\frac{s^2}{n}}}$$

g.l. = n -1



Análise Bivariada

COMPARAÇÃO DE PROPORÇÕES



Grupo A **480/500**

Grupo B **400/500**

1. $H_0 : p_A = p_B$
2. $H_0 : p_A \leq p_B$
3. $H_0 : p_A \geq p_B$

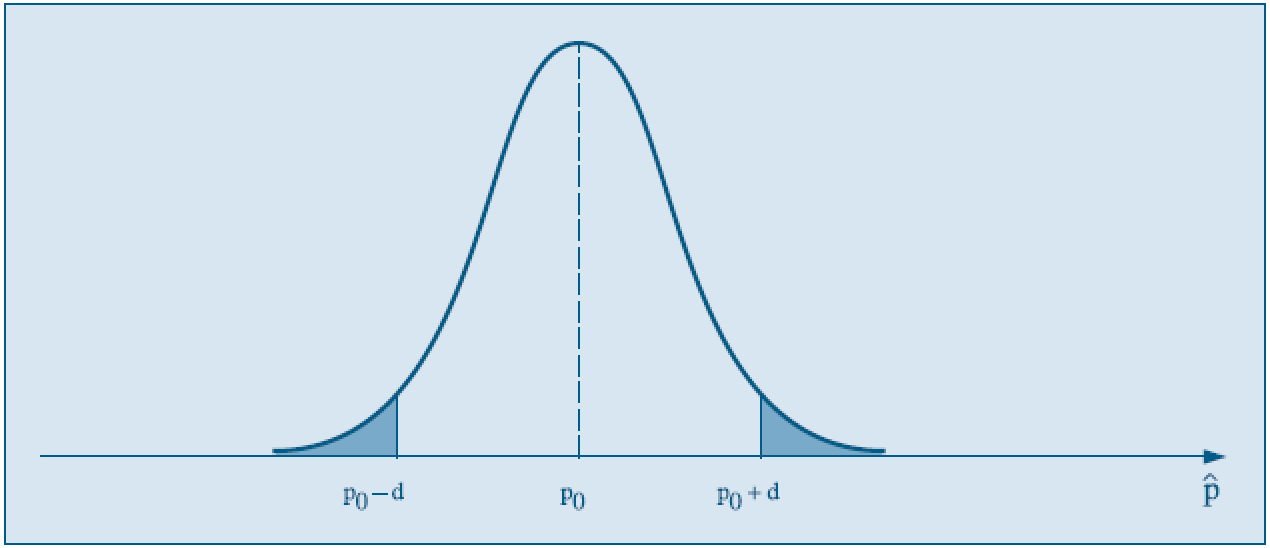
1. $H_a : p_A \neq p_B$
2. $H_a : p_A > p_B$
3. $H_a : p_A < p_B$

bicaudal

unicaudal

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

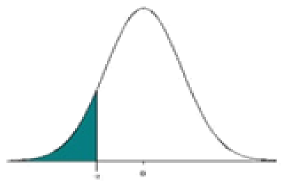
Figura 12.6: Região crítica para o teste $H_0 : p = p_0$ vs $H_1 : p \neq p_0$.



$$z = \frac{p_A - p_B}{\sqrt{pq/n_A + pq/n_B}}$$

Table of Standard Normal Probabilities for Negative Z-scores

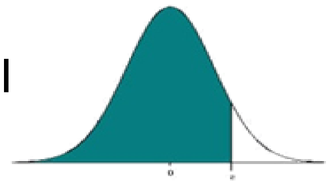
Tabela I



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores

Tabela II



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

```
prop.test(x, n,
          p = NULL,
          alternative = "two.sided“,
          correct = TRUE)
```

x: Vetor - Contagem do número de sucessos
n: Vetor – Contagem do número de tentativas
alternative: Caracter – especificando a hipótese alternativa
correct: Lógico – Se correção de Yates dve ser realizada

```
> prop.test(x = c(490,400),
+           n = c(500, 500))

	2-sample test for equality of proportions with continuity correction

data:  c(490, 400) out of c(500, 500)
X-squared = 80.909, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1408536 0.2191464
sample estimates:
prop 1 prop 2 
  0.98   0.80 

> 
```

→ Tamanho amostral

$$n = \frac{Z_{gc}^{2} \cdot p \cdot q}{e^2}$$

→ Z_{gc}: Valor de z relativo ao grau de confiança (95% = 1.96)

→ e: erro absoluto (“para mais ou para meno”)

→ p: probabilidade do evento

→ q: probabilidade do não evento (1-p)

→ Tamanho amostral

Interpretação prática:

Nós estamos 100(1-α)% confiantes que o intervalo de confiança contém o valor de μ

EXERCÍCIO

1. Os produtores de um programa de televisão pretendem modificá-lo se for assistido regularmente por menos de um quarto dos possuidores de televisão. Uma pesquisa encomendada a uma empresa especializada mostrou que, de 400 famílias entrevistadas, 80 assistem ao programa regularmente. Com base nos dados, qual deve ser a decisão dos produtores?

Temos uma população *P e queremos verificar se ela segue uma distribuição* especificada *P_0, isto é, queremos testar a hipótese H0 : $P = P_0$.*

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

Temos uma população *P e queremos verificar se ela segue uma distribuição* especificada *P_0, isto é, queremos testar a hipótese H0 : P = P_0.*

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Tabela 14.1: Resultados do lançamento de um dado 300 vezes.

Ocorrência (i)	1	2	3	4	5	6	Total
Freq. Observada (n_i)	43	49	56	45	66	41	300

Temos uma população *P e queremos verificar se ela segue uma distribuição* especificada *P_0, isto é, queremos testar a hipótese H0 : P = P_0.*

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

Tabela 14.1: Resultados do lançamento de um dado 300 vezes.

Ocorrência (i)	1	2	3	4	5	6	Total
Freq. Observada (n_i)	43	49	56	45	66	41	300
Freq. Esperada (n_i^*)							300

Temos uma população *P e queremos verificar se ela segue uma distribuição* especificada P_0*, isto é, queremos testar a hipótese H0 :* $P = P_0$.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Tabela 14.1: Resultados do lançamento de um dado 300 vezes.

Ocorrência (i)	1	2	3	4	5	6	Total
Freq. Observada (n_i)	43	49	56	45	66	41	300
Freq. Esperada (n^*_i)							300

$1/6 \rightarrow n^*_{ij}/300$

Temos uma população *P e queremos verificar se ela segue uma distribuição* especificada P_0*, isto é, queremos testar a hipótese* $H0 : P = P_0$.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{ij}^{*})^2}{n_{ij}^{*}}$$

Tabela 14.1: Resultados do lançamento de um dado 300 vezes.

Ocorrência (i)	1	2	3	4	5	6	Total
Freq. Observada (n_i)	43	49	56	45	66	41	300
Freq. Esperada (n_i^*)	50	50	50	50	50	50	300

Calcule o χ^2

EXERCÍCIOS

Cem estudantes foram divididos em duas classes de 50 cada e o objetivo era testar um novo método de ensinar Probabilidades. Uma classe recebeu um método tradicional e a outra, o novo método. Após o curso, foi pedido que os estudantes resolvessem um problema típico de Probabilidades. Os resultados foram os seguintes:

	Exercício correto	Exercício errado
Método convencional	33	17
Método novo	37	13

Há razões para acreditar que o novo método é superior?

Hardy-Weinberg

Genetic structure of parent population

phenotypes

genotypes	*AA*	*Aa*	*aa*
number of moths (total = 500)	320	160	20
genotype frequencies	$\frac{320}{500} = 0.64\ AA$	$\frac{160}{500} = 0.32\ Aa$	$\frac{20}{500} = 0.04\ aa$

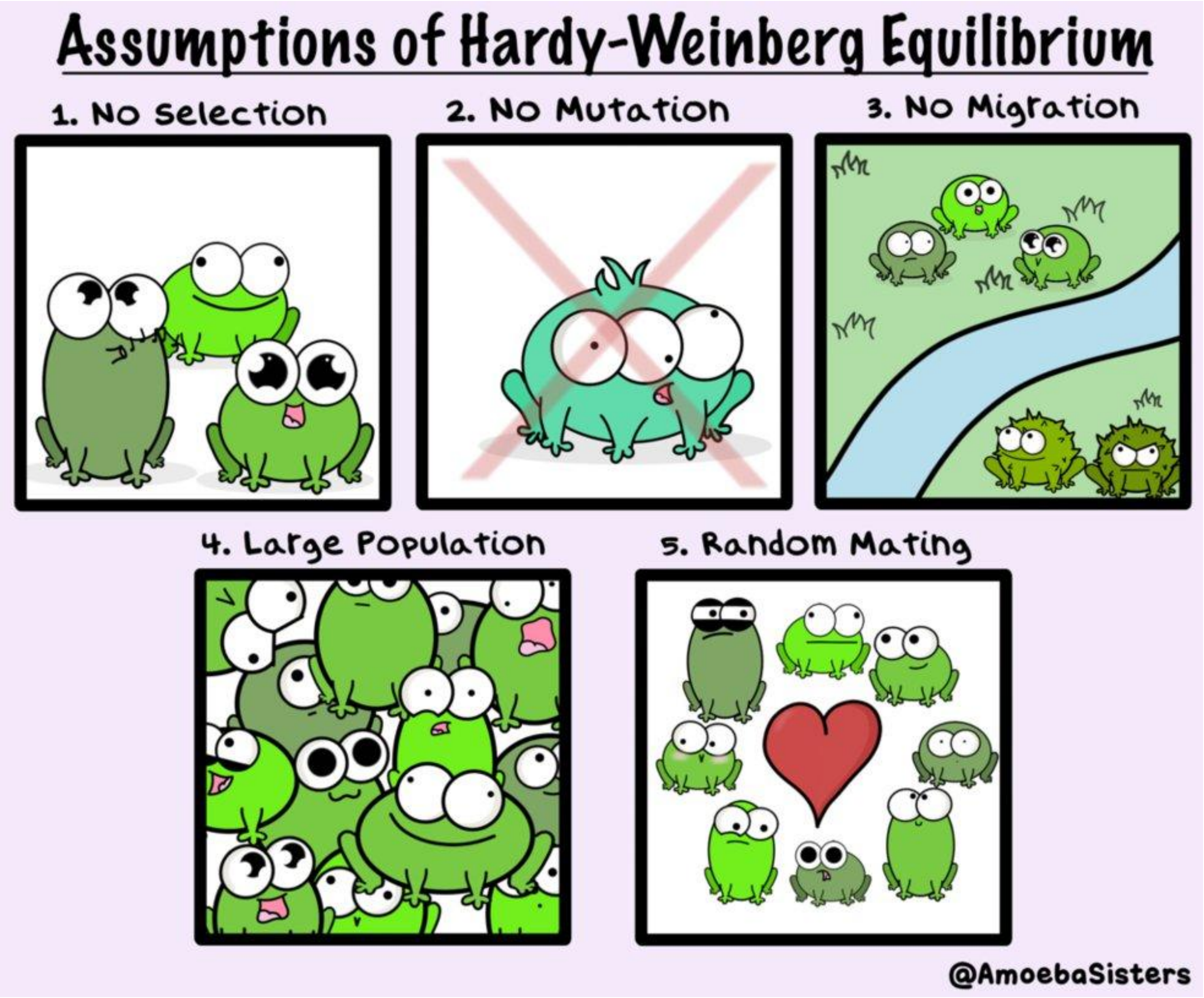
$\times 2$ | | $\times 2$

number of alleles in gene pool (total = 1000): 640 *A* | 160 *A* | 160 *a* | 40 *a*

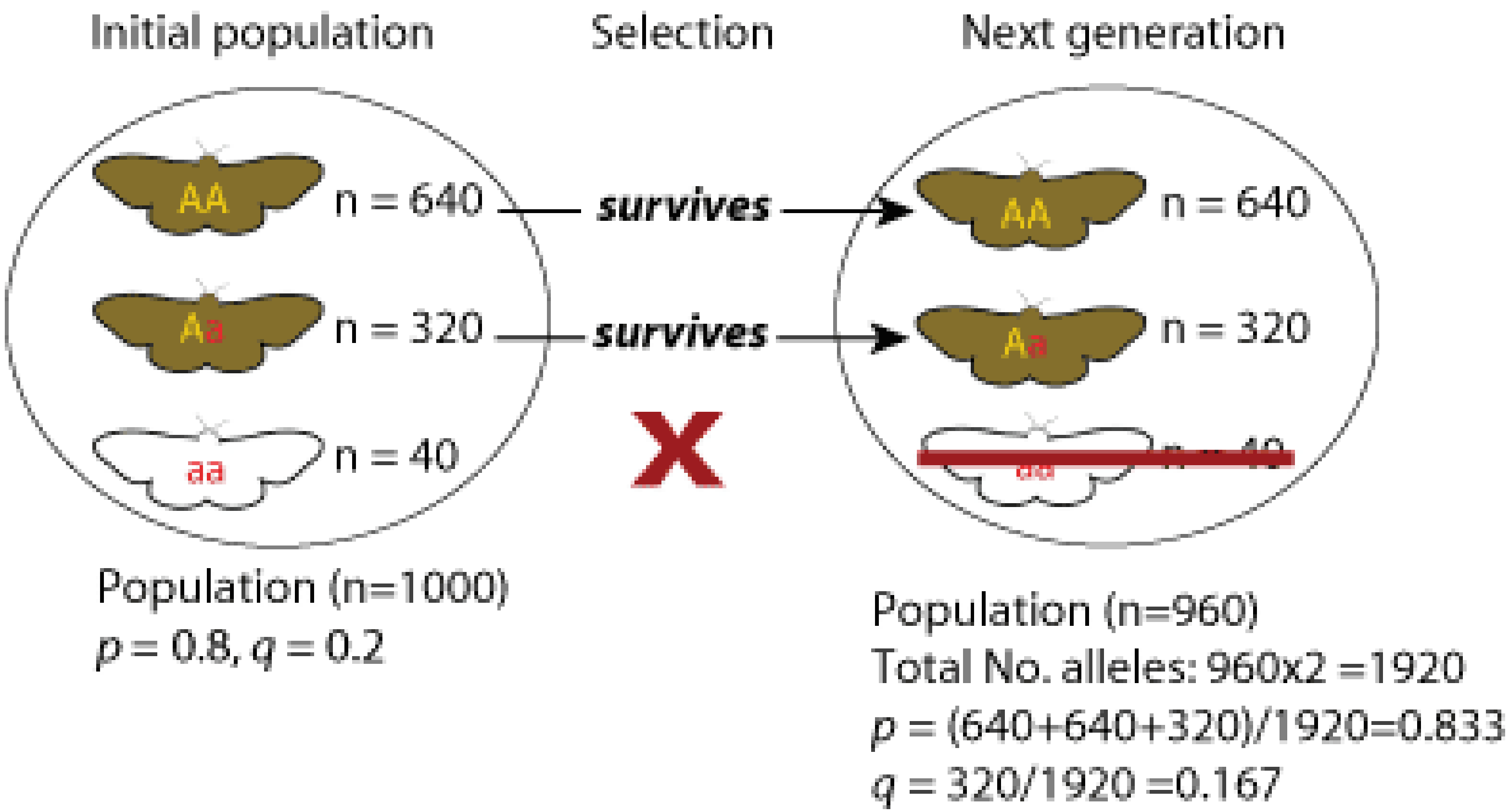
allele frequencies: $\frac{800}{1000} = 0.8\ A$ | $\frac{200}{1000} = 0.2\ a$

p = frequency of A = 0.8 | q = frequency of a = 0.2

Hardy-Weinberg



Hardy-Weinberg



Hardy-Weinberg

	AA	Aa	aa	Total
Observado	F_{AA}	F_{Aa}	F_{aa}	$N = F_{AA} + F_{Aa} + F_{aa}$
Esperado	$p^2 N$	$2pqN$	$q^2 N$	N
Contribuição para chi^2	$\frac{(F_{AA} - p^2 N)^2}{p^2 N}$	$\frac{(F_{Aa} - 2pqN)^2}{2pqN}$	$\frac{(F_{aa} - q^2 N)^2}{q^2 N}$	χ^2

3 frequências → g.l. = 2-1

EXERCÍCIOS

Determine se a população a seguir encontra-se em equilíbrio de Hardy- Weinberg

98 AA;
50 Aa;
20 aa

N = 168

Considere Qui-Quadrado crítico (associado à probabilidade de 5%) = 5,99

Probabilidade Condicional e Independência

Dois eventos, A e B de um mesmo espaço amostral são **independentes** quando a probabilidade de que eles ocorram simultaneamente, for igual ao produto de suas probabilidades individuais

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A)$$

$$P(A|B) = P(A)P(B)$$

A partir do resultado de um deles não é possível inferir nenhuma conclusão sobre o outro.

“Teste” de Independência

Se $P(A \cap B) \neq P(A)*P(B)$,

podemos considerar que existe dependência (ou associação) entre eventos.

Por exemplo:

P(H – hipertensos) = 23% = 0.23

P(M – hipertensas) = 18% = 0.18

P(casais – hipertensos) = 7.2%

→ Podemos considerar associação ou dependência?

EXERCÍCIOS

Realizou-se uma pesquisa com uma amostra de 400 frequentadores de um clube esportivo, sendo 150 mulheres e 250 homens, a fim de classificá-los de acordo com a modalidade esportiva preferida: vôlei, basquete ou tênis. Os dados coletados na pesquisa estão descritos na tabela a seguir:

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	75	25	50	150
HOMENS	40	150	60	250
TOTAL	115	175	110	400

a) Se uma pessoa é amostrada aleatoriamente, qual é a probabilidade desta pessoa ser uma mulher e praticar Vôlei?

EXERCÍCIOS

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	75	25	50	150
HOMENS	40	150	60	250
TOTAL	115	175	110	400

Construa uma tabela com as frequências esperadas entre modalidades e gênero

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES				150
HOMENS				250
TOTAL	115	175	110	400

EXERCÍCIOS

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	75	25	50	150
HOMENS	40	150	60	250
TOTAL	115	175	110	400

Construa uma tabela com as frequências esperadas entre modalidades e gênero

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	**Pmulher*Volei**	**Pmulher*basq**	**Pmulher*tenis**	150
HOMENS	**Phomem*Volei**	**Phomem*basq**	**Phomem*tenis**	250
TOTAL	115	175	110	400

EXERCÍCIOS

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	75	25	50	150
HOMENS	40	150	60	250
TOTAL	115	175	110	400

Construa uma tabela com as frequências esperadas entre modalidades e gênero

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	(150/400)*115	(150/400)*175	(150/400)*110	150
HOMENS	(250/400)*115	(250/400)*175	(250/400)*110	250
TOTAL	115	175	110	400

EXERCÍCIOS

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	75	25	50	150
HOMENS	40	150	60	250
TOTAL	115	175	110	400

Construa uma tabela com as frequências esperadas entre modalidades e gênero

GÊNERO	VÔLEI	BASQUETE	TÊNIS	TOTAL
MULHERES	**43.13**	**65.625**	**41.25**	150
HOMENS	**71.88**	**109.375**	**68.75**	250
TOTAL	115	175	110	400

EXERCÍCIOS

Calcule o valor do Qui-quadrado.

Considere o g.l. = *(linhas – 1)*(colunas – 1).*

$$\chi^2 = \sum \frac{(Obs - Esp)^2}{Esp}$$

EXERCÍCIOS

Uma pesquisa sobre a qualidade de certo produto foi realizada enviando-se questionários a donas-de-casa pelo correio. Aventando-se a possibilidade de que os respondentes voluntários tenham um particular viés de respostas, fizeram-se mais duas tentativas com os não-respondentes. Os resultados estão indicados abaixo. Você acha que existe relação entre a resposta e o número de tentativas?

Opinião sobre o produto	Nº de donas-de-casa		
	1ª tentativa	2ª tentativa	3ª tentativa
Excelente	62	36	12
Satisfatório	84	42	14
Insatisfatório	24	22	24

Teste de Homogeneidade

Verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações. Em outras palavras, em um teste de Chi Quadrado de homogeneidade podemos testar a afirmação de que diferentes populações têm a mesma proporção de indivíduos com alguma característica.

Teste de Independência

O teste de Chi Quadrado de independência é semelhante ao teste de Chi Quadrado de aderência, mas considera uma "lei oriunda da própria tabela de dados experimentais" a fim de avaliar se há ou não dependência entre duas variáveis. Quanto maior a dependência entre as duas variáveis, maior será o valor de χ2. Quando as duas variáveis são independentes, o valor de χ2 tende a zero

Diferenças

Teste de homogeneidade: selecionamos uma amostra de elementos de cada uma das ***r* subpopulações e distribuímos os elementos de cada uma dessas amostras segundo *x* categorias.**

Teste de independência: distribuímos uma amostra de **N elementos de ”uma” população segundo as categorias da variável A e as categorias da variável B.**

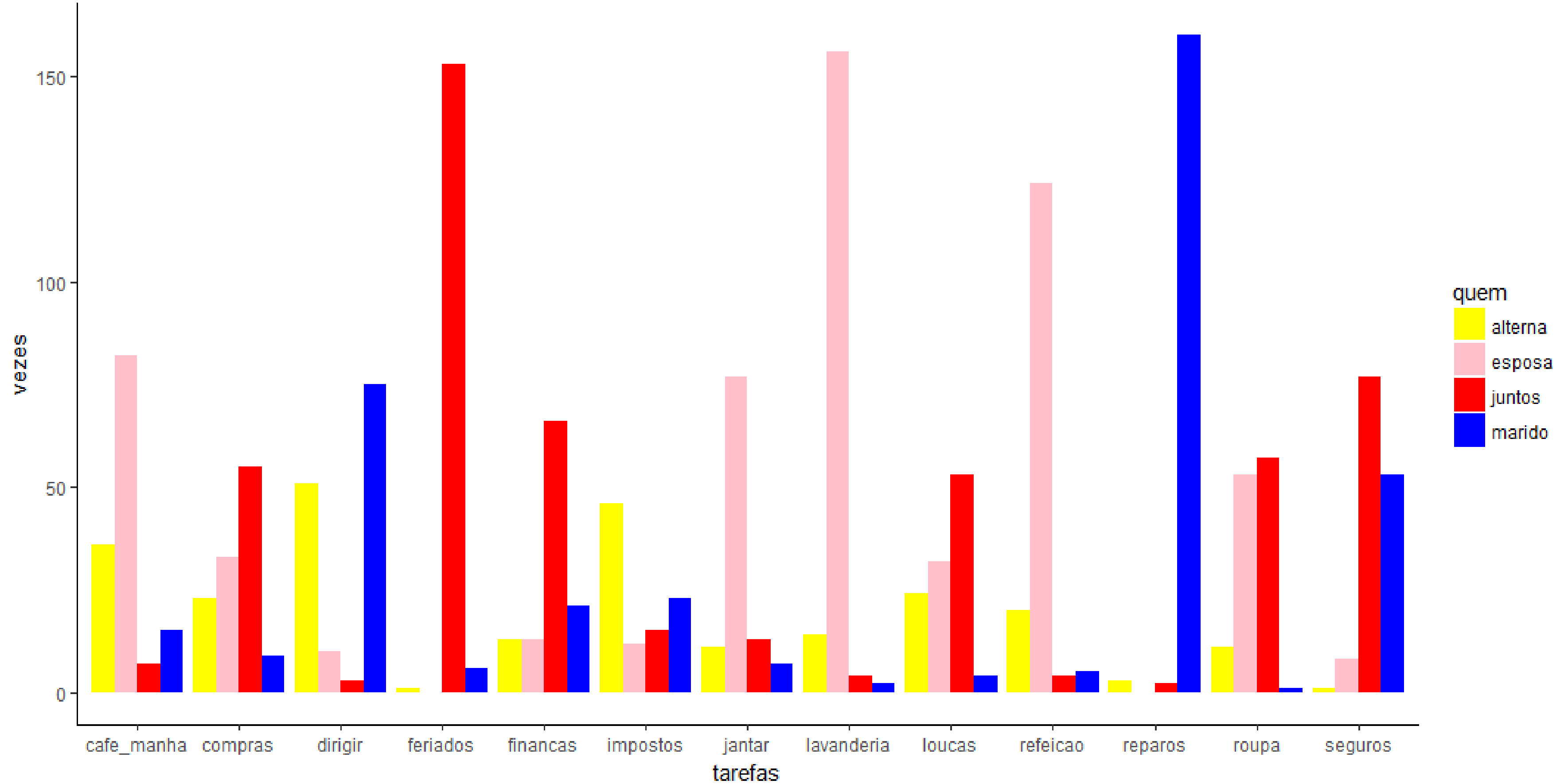
EXERCÍCIOS

Um sociólogo afirma que a distribuição de idades dos moradores de certa cidade é diferente do que era 10 anos antes. Você seleciona aleatoriamente 400 moradores e registra a idade de cada um deles. Os resultados são registrados na tabela abaixo. Pode-se afirmar, com alfa igual a 5%, que a distribuição de idades foi alterada nesses 10 anos? E com alfa igual a 1%?

Idade	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
Anterior	16%	20%	8%	14%	15%	12%	10%	5%
Pesquisa	76	84	30	60	54	40	42	14

EXERCÍCIOS

```
read.csv("tarefas.csv")
tarefas_long<-gather(tarefas, key=quem, value=vezes)
tarefas_long$tarefas<-row.names(tarefas)
ggplot(tarefas_long, aes(x=tarefas, y=vezes, fill=quem))+
  geom_col(position = 'dodge')+ theme_classic()+
  scale_fill_manual(values=c('yellow', 'pink', 'red', 'blue'))
```



Dados: http://www.sthda.com/sthda/RDoc/data/housetasks.txt

PRESSUPOSTOS

1º Pelo menos 5 observações por casela
2º Menos de 20% das caselas com ZERO
3º N total mínimo = 5 x N de caselas

ALTERNATIVAS

→ Teste Exato de Fisher

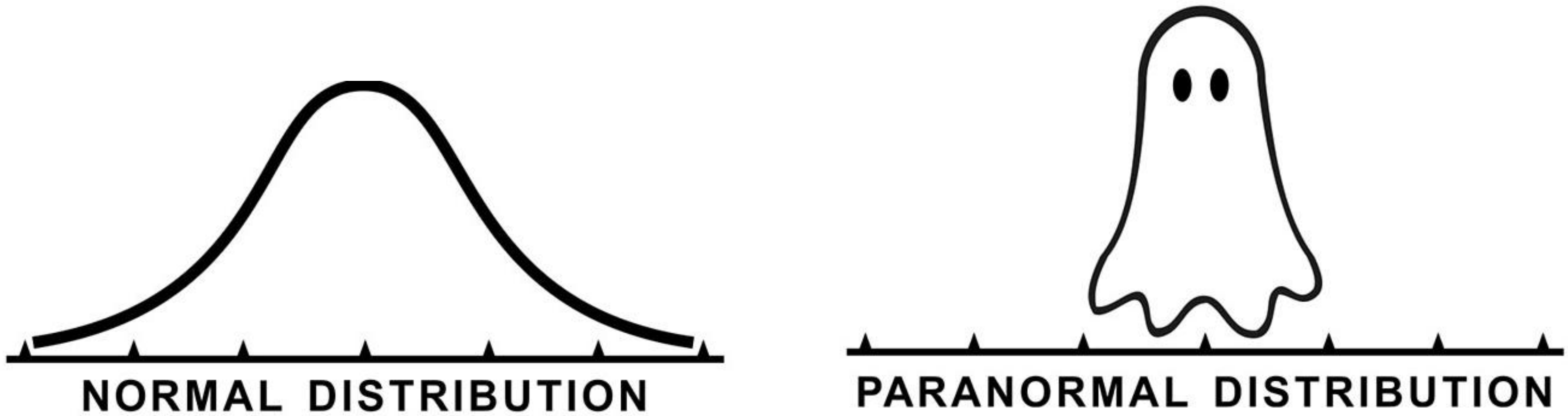
Utiliza tabelas 2x2 e N total menor que 20 (aceita ZERO)

→ Teste de McNemar

Quando existe medidas "antes" e "depois".

→ Teste de Mantel-Haenszel

Quando existe medidas com interferência de alguma variável associada: ***Confounding***



Análise Bivariada

TESTES NÃO-PARAMÉTRICOS
(*OUTROS*)

Teste dos Sinais Teste Mann-Witney Teste de Wilcoxon